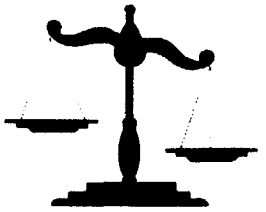
Non-Normality

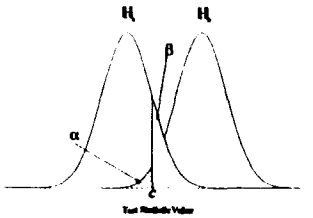


- BellSouth and Original LCUG about the same at levels of testing BST advocates

Small Sample Sizes

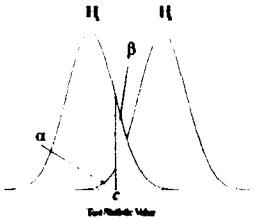
- At aggregate level that BST advocates, the problem of small sample sizes is not an issue.

Efficiency



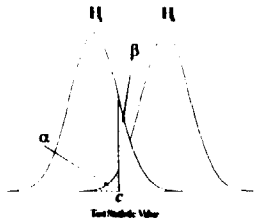
- Relative Power of Tests
- Other Considerations

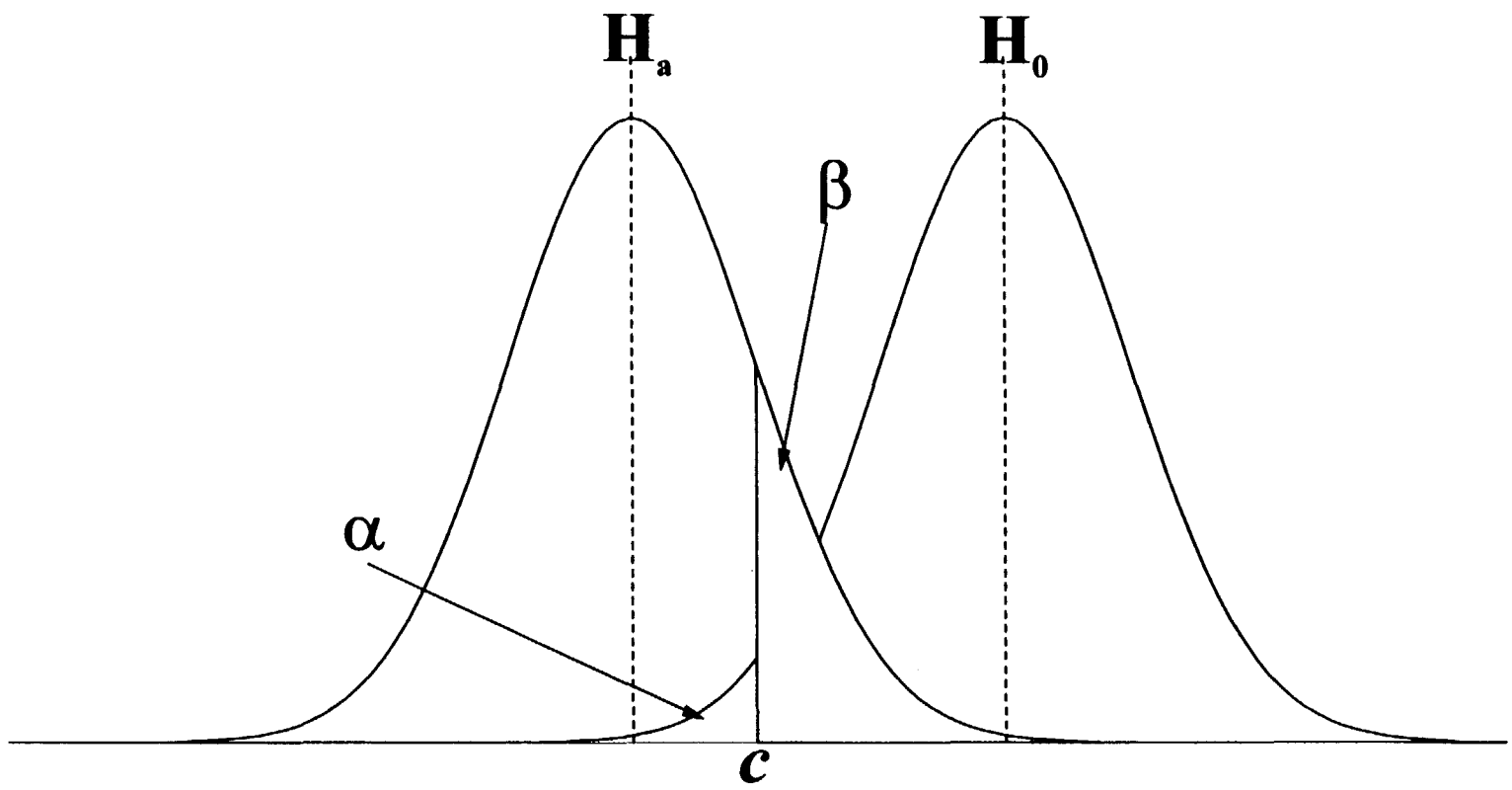
Relative Power of Tests



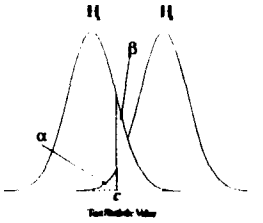
- Generally appear to be about the same when independence holds.
- BST approach can explicitly balance risk. This is unclear for most recent LCUG.

Balancing Error





Other Considerations



- BellSouth
  - A complete operating system, fully responsive
  - Handles dependencies within/across measures
  - Computer intensive
- LCUG
  - Still under development
  - Does not handle dependency satisfactorily
  - More computer intensive

Estimating Variance

Advantage of Jackknife Method

- More robust to model misspecification
- Handles dependence within data

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Aggregating Data

FCC Q - 5, 6

- 3 Comparisons
- Illustrating Validity
- Middle Ground?

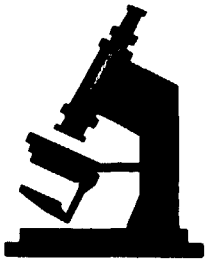
3 Comparisons

- Aggregated Adjusted Data (BST)
- Aggregated Unadjusted Data (original LCUG)
- Disaggregated Data (most recent LCUG)

Illustrating Validity

- Observational study ⟶ worry about **bias**
- To reduce **bias**, control for confounding factors as in a designed experiment -- Time, location, etc.

Middle Ground?



- Appropriate "Middle Ground" would change from month to month
- Therefore, not feasible or consistent with black box / production mode

Dependency

- Effect of Dependency on Jackknife Method
- Comparison of Effects of Dependency
- Measuring Dependency
- Effect of Dependency on Type I Error

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Jackknife Method

- Captures covariance component when it is not zero
- When covariance is zero, handles estimate in same way as other test
- Reduces covariance contribution across wire centers.
  - This is possible when there is correlation between subclass differences within a wire center, but no correlation between subclass differences from different wire centers.

Basic Theory

The estimate of the difference, $\hat{D}$ can be written as a linear combination of the form

$$\sum_{k=1}^{N} c_k d_k$$

where the c_k are constants and the d_k is the ILEC - CLEC mean difference of subclass k.

Basic Theory

The variance of a linear combination such as this can be calculated as

$$Var(\sum_k c_k d_k) = \sum_k c_k^2 Var(d_k) + \sum_k \sum_{j \neq k} c_k c_j Cov(d_k, d_j)$$

where $Cov(d_k,d_j)$ is the covariance between the two quantities.

Key Difference

LCUG Approach:

$$Var(\sum_k c_k d_k) = \sum_k c_k^2 Var(d_k) + \left(\sum_k \sum_{j \neq k} c_k c_j Cov(d_k, d_j)\right)$$

Treats this term = 0

Key Difference

- Ignoring covariance term can lead to double counting.

Effects Of Dependence

- BST Jackknife Method
  - No effect on Type I Error
  - Properly reflects reduced power, therefore increasing Type II Error
- LCUG
  - Type I Error unfairly inflated
  - Type II Error no longer in balance

Measuring Dependence

- Statistical Methods
- General Physical Relationships
- Individual Events
- Covariance Matrix
  - Level of Aggregation?

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Cluster Effect (CEFF)

$$CEFF = \frac{\text{Jackknife Variance}}{\text{Simple Random Sample Variance}}$$

Effective Sample Size

Effective Sample Size:

$$n_b^* = \frac{n_b}{CEFF} \qquad n_c^* = \frac{n_c}{CEFF}$$

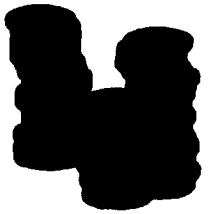
n_b^* and n_c^* Can Be Used In Formula For Simple Random Samples To Calculate:

- Type II Error
- Balancing Critical Value

Effect of Dependence on Type I Error

- BST
  - No effect
- LCUG
  - Increases

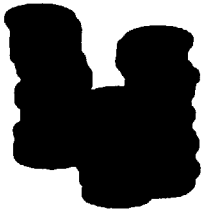
Statistical vs. Competitive Significance of Results



- Statistical Difference = Discrimination?
- Economic Impact
- Determining "Threshold Difference"

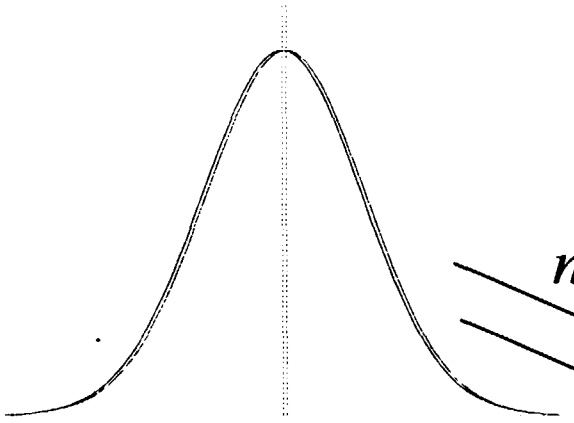
EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Statistical Difference = Discrimination?



2 Normal Distributions

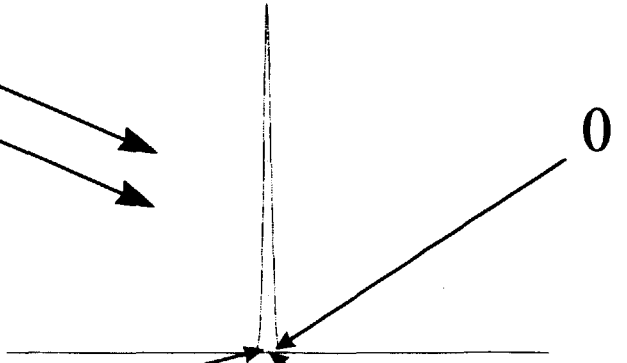
With large enough sample sizes, even tiny differences can be statistically significant.



$\mu_1 = 1$ $\mu_2 = 1.05$

n_1, n_2 large

Distribution of $\bar{x}_1 - \bar{x}_2$



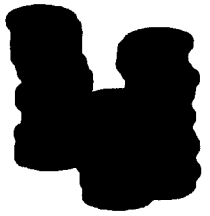
0

$\mu_{\bar{x}_1 - \bar{x}_2} = -0.05$

significant gap $> 2\sigma_{\bar{x}_1 - \bar{x}_2}$

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Does *Statistical Significance* Imply *Practical Significance* ???

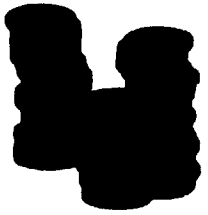


"*Remember also that a significant* t *value is evidence only that the population* means *differ. Popular accounts are sometimes written as if a significant* t *implies that every member of population 1 is superior to every member of population 2.... In fact, the two populations usually overlap substantially even though* t *is significant.*"

(Snedecor and Cochran, *Statistical Methods*)

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Does *Statistical Significance* Imply *Practical Significance* ???



- With very large sample sizes, even small differences can be statistically significant
- When does statistically significant differences in means imply discrimination in service?
- When does a difference in means have an economic impact for CLECs?
- How to determine a practical significance threshold for performance measures?

Appendix

Test Statistics

Notation:

n_1 = the number of BST cases

n_{1j} = the number of BST cases in subclass j

x_{1i} = the value of the performance measure for the i^{th} BST observation

$\bar{x}_1$ = the mean of the BST observations

$\bar{x}_{1j}$ = the mean of the BST observations in subclass j

$$\bar{x}_{1w} = \frac{1}{n_2}\sum_j n_{2j}\bar{x}_{1j} = \frac{\sum_j w_{1j}\sum_{i=1}^{n_{1j}} x_{1i}}{\sum_j w_{1j}n_{1j}} \quad \text{where } w_{1j} = \frac{n_{2j}}{n_{1j}}$$

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2}{(n_1 - 1)}$$

$$s_{1w}^2 = \frac{\sum_j w_{1j}\sum_{i=1}^{n_{1j}}(x_{1i} - \bar{x}_{1w})^2}{\sum_j w_{1j}n_{1j} - 1}$$

$$c_1 = \frac{\sum_j w_{1j}^2 n_{1j}}{(\sum_j w_{1j}n_{1j})^2}$$

$$s_{1w2}^2 = \text{var}(\bar{x}_{1w} - \bar{x}_2) = \frac{\sum_j n_{2j}^1(w_{1j}^1 + 1)s_{1j1}^2 + n_{2j}^2(w_{1j}^2 + 1)s_{1j2}^2}{(n_2)^2} \text{ and}$$

s_{1j1}^2 and s_{1j2}^2 are the sample variances of ILEC observations in subclass 1 and 2 in wire center *j*.

Similar notation using the subscript 2 is used to denote the values for the CLEC cases, that is

n_2 = the number of CLEC cases, etc.

Table 1: Test Statistics

Test	Formula
Modified Z	$\frac{\bar{x}_1 - \bar{x}_2}{s_1\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$
Adj. Modified Z 1	$\frac{\bar{x}_{1w} - \bar{x}_2}{s_{1w}\sqrt{c_1+\frac{1}{n_2}}}$
Adj. Modified Z 2	$\frac{\bar{x}_{1w} - \bar{x}_2}{s_{1w2}}$
Jackknife	$\frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}}$
Adj. Jackknife 1	$\frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}} \cdot \frac{\sqrt{c_1 s_{1w}^2 + \frac{s_2^2}{n_2}}}{s_{1w}\sqrt{c_1+\frac{1}{n_2}}}$
Adj. Jackknife 2	$\frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}} \cdot \sqrt{\frac{\sum_j [(w_{1j}^1)^2 n_{1j}^1 s_{1j1}^2 + (w_{1j}^2)^2 n_{1j}^2 s_{1j2}^2] + \sum_j [n_{2j}^1 s_{2j1}^2 + n_{2j}^2 s_{2j2}^2]}{\sum_j [n_{2j}^1 (w_{1j}^1+1) s_{1j1}^2 + n_{2j}^2 (w_{1j}^2+1) s_{1j2}^2]}}$

Simulation Procedure

The simulation was carried out as follows.

1. Generate ILEC and CLEC sample sizes as follows. Draw n, the sum of ILEC and CLEC sizes, from a Poisson distribution with λ=29120. Split n into n_1 and n_2, the ILEC size and the CLEC size, by generating p from Uniform(0.025, 0.075), n_2 from Binomial(n_2, p) and n_1=n - n_2.

2. Generate ILEC and CLEC wire center sizes. For ILEC, draw the wire center sizes n_{1j}, j=1, …, 240, from a Multinomial (n_1, 240, p_s), where the probability vector p_s is generated from a Dirichelet distribution. Do the same thing to generate the CLEC wire center sizes n_{2j}, j=1, …,240. If one of the n_{2j} is 0, then the corresponding wire center is excluded from further analysis.

3. Generate the ILEC and CLEC observations within each wire center from multivariate normal. For ILEC, draw the observations from a multivariate normal with mean vector **0** and correlation matrix

$$\begin{bmatrix} 1 & \rho_j & \cdots & \rho_j & \rho_j \\ \rho_j & 1 & \cdots & \rho_j & \rho_j \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_j & \rho_j & \cdots & 1 & \rho_j \\ \rho_j & \rho_j & \cdots & \rho_j & 1 \end{bmatrix}$$

where ρ_j (j = 1,...,240) is from a Uniform(a, b) where no correlation (independence) is given by $a = b = 0$, medium correlation is given by $a = 0.1$ and $b = 0.15$, and high correlation is given by $a = 0.25$ and $b = 0.5$. The observations from different wire centers are independent of each other. Generate the CLEC sample using the same method. The resulting draws are correlated if they are from the same wire center, and independent if they are from different wire centers.

4. Split the observations within each wire center into two subclasses. For ILEC observations, draw the splitting probability p_{sp} from Uniform(0.65, 0.75); generate the first subclass size n_{1j}^1 from Binomial(n_{1j}, p_{sp}), where n_{1j} is the j^{th} ILEC wire center size; and calculate the second subclass size n_{1j}^2 using $n_{1j}^2 = n_{1j} - n_{1j}^1$. The first n_{1j}^1 draws of the ILEC observations in wire center j is the first subclass for wire center j and the rest is the second subclass. Split the CLEC sample using the similar method. n_{2j}^1 and n_{2j}^2 are the first and second subclass size of the CLEC for wire center j. Since there are three possible outcomes of $n_{1j}^1, n_{2j}^1, n_{1j}^2$ and n_{2j}^2 combinations, which subclass to use in the test statistics calculation depends upon the actual $n_{1j}^1, n_{2j}^1, n_{1j}^2$ and n_{2j}^2 values.

   a) If $n_{1j}^1 > 0$, $n_{2j}^1 > 0$, $n_{1j}^2 > 0$ and $n_{2j}^2 > 0$, then the observations in both subclasses of ILEC and CLEC are included in the calculation.
   b) If $n_{1j}^1 > 0$, $n_{2j}^1 > 0$ and either $n_{1j}^2 = 0$ or $n_{2j}^2 = 0$, then only the observations in the first subclass are used in the calculation.
   c) If either $n_{1j}^1 = 0$ or $n_{2j}^1 = 0$ and $n_{1j}^2 > 0$ and $n_{2j}^2 > 0$, then only the observations in the second subclass are included in the calculation.

   Denote the actual ILEC and CLEC sample size again as n_1 and n_2 for ease of notation.

5. Generate wire center mean effects, m_j, from Beta(2,3) and standard deviation effects, t_j, from a Uniform(1, 1.2). Generate the subclass 1 mean effect, v_1, from a Uniform(0, 1.5), and standard deviation effect, w_1, from a Uniform(1, 1.05). Generate the subclass 2 mean effect, v_2, from a Uniform(1, 5), and standard deviation effect, w_2, from a Uniform(1.05, 2). Rescale and shift each observation generated in (3) by

amounts corresponding to the wire center and subclass the observation is in. For modeling discrimination against CLECs, include scale and shift discrimination factors. That is,

$$X_{jk} = \sqrt{r}\, t_j w_k X'_{jk} + u_j + v_k + d t_j w_k,$$

where X'_{jk} is a multivariate normal observation in wire center j, subclass k generated in step (3), d is a mean discrimination factor, and r is a variance discrimination factor. For ILEC observation, d = 0, and r = 1. For CLEC observations, d > 0 and/or r > 1 models discrimination.

6. Calculate the test statistics in Table 1. For the Jackknife test statistics calculation, sort the wire centers according to ILEC wire center sizes, group every 30 wire centers sequentially to form 8 groups, permute the wire centers within each of the 8 groups to reduce bias, and select one wire center from each group to form a replicate. We have a total of 30 replicates. Calculate an estimator $\hat{D}$ from the full data set using

$$\hat{D} = \frac{\sum_j [n^1_{2j}(\bar{x}^1_{1j} - \bar{x}^1_{2j}) + n^2_{2j}(\bar{x}^2_{1j} - \bar{x}^2_{2j})]}{n_2},$$

where $\bar{x}^1_{1j}$ and $\bar{x}^2_{1j}$ are the first and second subclass mean of ILEC in wire center j and $\bar{x}^1_{2j}$ and $\bar{x}^2_{2j}$ are the first and second subclass mean of CLEC in wire center j. Let $\hat{D}_{(g)}$ denote the estimator of the same functional form as $\hat{D}$ but calculated from the observations removing the g^{th} replicate. Define the g^{th} pseudo-value as

$$\hat{D}_g = 30 \cdot \hat{D} - 29 \cdot \hat{D}_{(g)}.$$

There are total 30 pseudo-values. Calculate the Jackknife statistics using

$$t = \frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}},$$

where $\hat{\bar{D}} = \frac{1}{30}\sum_{g=1}^{30} \hat{D}_g$ and $v(\hat{\bar{D}}) = \frac{1}{30(30-1)}\sum_{g=1}^{30}(\hat{D}_g - \hat{\bar{D}})^2$. Calculate the adjusted Jackknife 1 test statistics using

$$t = \frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}} * \frac{\sqrt{c_1 s_{1w}^2 + \frac{s_2^2}{n_2}}}{s_{1w}\sqrt{c_1 + \frac{1}{n_2}}},$$

where s_2 is the regular standard error of the CLEC observations. Compute the adjusted Jackknife 2 test statistics as follows.

$$t = \frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}} * \sqrt{\frac{\sum_j [(w_{1j}^1)^2 n_{1j}^1 s_{1j1}^2 + (w_{1j}^2)^2 n_{1j}^2 s_{1j2}^2] + \sum_j [n_{2j}^1 s_{2j1}^2 + n_{2j}^2 s_{2j2}^2]}{\sum_j [n_{2j}^1 (w_{1j}^1 + 1) s_{1j1}^2 + n_{2j}^2 (w_{1j}^2 + 1) s_{1j2}^2]}},$$

where s_{2j1} and s_{2j2} are the sample variances of CLEC observations in subclass 1 and 2 in wire center j, respectively.

9. Compare all the test statistics with the critical value -1.65.

Repeat the above procedure 1000 times to estimate the type I or type II error of the corresponding test.

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

BellSouth Test

- Prepare Data for Statistical Analysis
- Add Weights to Observations
- Generate Statistics
- Generate Replicates
- Perform Jackknife Analysis
- Interpret Results

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Jackknife Estimate and Test Statistic

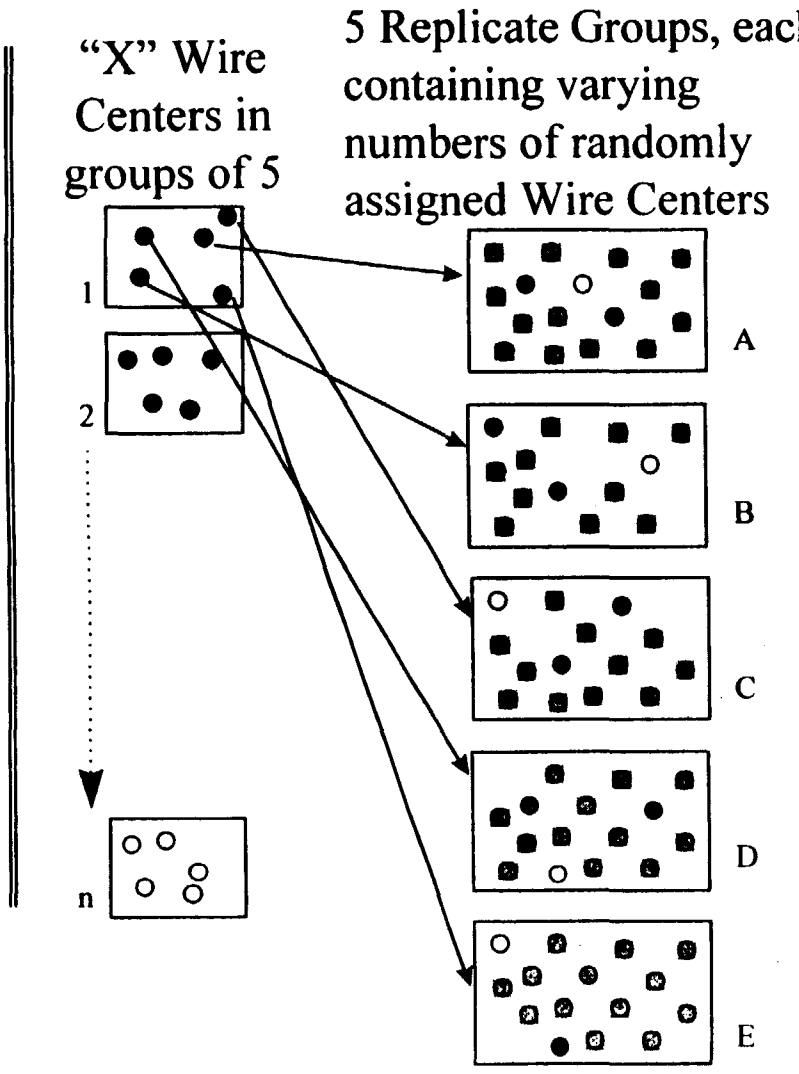
- Reduces Bias
- Estimate Variance of

$$\hat{D} = \frac{1}{n_c}\sum_j n_{cj}(\bar{x}_{bj} - \bar{x}_{cj})$$

- The Observations Are Partitioned Into G Groups or Replicates. g=1,2,… G

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Jackknife Method



Statistical processes are performed on each of these groups of replicates, dropping a different replicate each time.

$$B, C, D, E \rightarrow \hat{D}_{(A)}$$
$$A, C, D, E \rightarrow \hat{D}_{(B)}$$
$$A, B, D, E \rightarrow \hat{D}_{(C)}$$
$$A, B, C, E \rightarrow \hat{D}_{(D)}$$
$$A, B, C, D \rightarrow \hat{D}_{(E)}$$

Jackknife Estimate

- $\hat{D}_{(g)}$ is calculated similar to $\hat{D}$ except remove the gth group.
- G Pseudo Values: $\hat{D}_g = G \cdot \hat{D} - (G-1) \cdot \hat{D}_{(g)}$

Jackknife Estimate

- Mean of the Pseudo Values: $\hat{\bar{D}} = \frac{1}{G}\sum_{g=1}^{G}\hat{D}_g$

Jackknife Estimate

- Variance of $\hat{\bar{D}}$: $v(\hat{\bar{D}}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (\hat{D}_g - \hat{\bar{D}})^2$

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

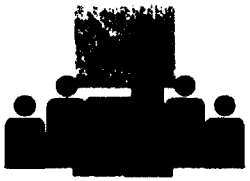
Jackknife Test Statistic

Test Statistic: $$\frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}}$$

t distribution with G-1 degrees of freedom.

EY econ
Economics Consulting & Quantitative Analysis
EY/Econ-STAT

Jackknife Test Statistic (Continued)



- The statistic

$$t = \frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}}$$

is distributed approximately as a Student's t with G-1 degrees of freedom. This is the test statistic recorded on the Decision page as the JACK test.

Jackknife Test Statistic (Continued)

- The adjusted jackknife, referred to on the Decision Page as JACK ADJ, is this t-statistic multiplied by the adjustment factor for unequal variances.

$$adj.fact = \sqrt{\frac{\frac{s_b^2}{n_b} + \frac{s_c^2}{n_c}}{s_b^2\left(\frac{\sum w_j^2}{\left(\sum w_j\right)^2} + \frac{1}{n_c}\right)}}$$

Statistical Procedure Off-Line Session
Consensus/Open Issues

Issue No.	Issue	Position
1	Comparing like-to-like	**Agreement:** In order to assure that like-to-like comparisons are made, the performance measure data must be disaggregated to a very deep level. This includes wire center and time of month, as well as SQM disaggregation levels defined by the Louisiana Public Service Commission.[D]
2	Performance measure test statistic	**Agreement:** Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists.
3	Methodology for obtaining the test statistic	**Dr. Mallows/LCUG**: In each cell, construct an indicator that is sensitive to absence of parity.. Make appropriate allowance for what would be the effect of random variation, assuming parity holds. The aggregate statistic should not allow consistent violations in any cell to go undetected. **BellSouth**: The overall service process is what defines parity. Testing measures at an aggregate level is sufficient to determine favoritism. Random failures at deeply disaggregated levels may exists but should not be overemphasized. SQM level disaggregation reports will be available to explore the data.
4	Type I and Type II errors	**Agreement:** The probability of a Type I error, concluding BellSouth favoritism exists when it does not, should be balanced with the probability of a type II error, concluding there is no BellSouth favoritism when there is. The balance of these two probabilities depends on 1. The effective number of BellSouth observations 2. The effective number of CLEC observations 3. The size of a specific alternative hypothesis, e.g., the CLEC mean value is larger than the BellSouth mean value by ten percent of a BellSouth standard deviation Using this information, a critical value for the test, or decision rule, is determined. This rule may be different for each performance measure in interest, and may also change over the months. However, a system can be devised to make this all transparent to the commission.

[D] Louisiana Public Service Commission Docket No. U-22252-Subdocket C, In Re: BellSouth Telecommunications Inc., Service Quality Performance Measurements, April 19, 1998 Order. Except that for provisioning measures order type was also included since there is a noticeable difference in their distributions.
Meeting between Dr. Colin Mallows and Dr. Fritz Scheuren on April 7, 1999, supplemented by later discussions.

Issue No.	Issue	Position
4a	Type I and Type II errors	**Dr. Mallows/LCUG**: We do not agree that the following BellSouth alternative is either feasible (since it requires the parties to agree on what constitutes a material difference), or fair (since it uses a test procedure at a level (2 1/2%) that is biased in favor of BellSouth for all sample sizes below 1000). **BellSouth**: If the balancing procedure described in Issue Number 4 is determined to be unworkable, then a feasible alternative is to define the size of a difference between mean values which has no business impact (a rule of materiality). Any actual difference less than this will be considered insignificant. Differences greater than the materiality standard would be judged to be significant based on a statistical testing procedure. This should be a five percent (5%) significance level, two-sided test (a two and one half percent (2.5%) significance level, one-sided test).
5	Statistical paradigm	**Agreement:** The system must be developed so that it can be put into production (black box). Two statistical paradigms are possible for examining the performance measure data. In the exploratory paradigm, data are examined and methodology is developed that is consistent with what is found. In a production paradigm a methodology is decided upon before data exploration. While the exploratory paradigm provides protection against using erroneous data it requires a great deal of lead time and is unsuitable for timely monthly performance measure testing. A production paradigm will not only promptly produce overall test results but will also provide documentation that can be used to explore the data after the test results are released.
6	Trimming	**Agreement:** Trimming is needed but finding a robust rule that can be used in a production setting is difficult. Trimming of extreme observations from BellSouth and CLEC distributions is needed in order to ensure that a fair comparison is made between performance measures. However, trimmed observations should not simply be discarded. They need to be examined and possibly used in the final decision making process. Under a production paradigm this is very hard to do. Additionally, each performance measure may need to use a different trimming rule.
7	Independence of performance measure tests	**Agreement:** Correlation between the performance measures must be accounted for in aggregation over performance measures.